Automatic Classification of Cancer Notifiable Death Certificates

Luke Butt[1], Guido Zuccon[1], Anthony Nguyen[1],
Anton Bergheim[2], Narelle Grayson[2]

[1] The Australian e-Health Research Centre, Brisbane, Queensland, Australia;
[2] Cancer Institute NSW, Alexandria, New South Wales, Australia.
{luke.butt, guido.zuccon, anthony.nguyen}@csiro.au
{anton.bergheim, narelle.grayson}@cancerinstitute.org.au

Abstract. The timely notification of cancer cases is crucial for cancer monitoring and prevention. However, the abstraction and classification of cancer from the free-text of pathology reports and other relevant documents, such as death certificates, are complex and time-consuming activities. In this paper we investigate approaches for the automatic detection of cases where the cause of death is a notifiable cancer from free-text death certificates supplied to Cancer Registries. A number of machine learning classifiers were investigated. A large set of features were also extracted using natural language techniques and the Medtex toolkit; features include stemmed words, bi-grams, and concepts from the SNOMED CT medical terminology. The investigated approaches were found to be very effective in identifying death certificates where the cause of death was a notifiable cancer. Best performance was achieved by a Support Vector Machine (SVM) classifier with an overall F-measure of 0.9647 when evaluated on a set of 5,000 free-text death certificates. This classifier considers as features stemmed token bigrams and information from SNOMED CT concepts filtered by morphological abnormalities and disorders. However, our analysis shows that it is the selection of features that most influences the performance of the classifiers rather than the type of classifier or the feature weighting schema. Specifically, we found that stemmed token bigrams with or without SNOMED CT concepts are the most effective feature. In addition, the combination of token bigrams and SNOMED CT information was found to yield the best overall performance.

Keywords: death certificates, Cancer Registry, cancer monitoring and reporting, machine learning, natural language processing, SNOMED CT

1 Introduction

Cancer notification and reporting is an important and fundamental process for providing an accurate picture of the impact of cancer, the nature and extent of cancer, and to direct research efforts for the cure of cancer. Cancer Registries collect and interpret data from a large number of sources, helping to improve cancer prevention and control, as well as treatments and survival rates for patients with cancer.

The manual coding of documents, such as pathology reports and death certificates, with respect to notifiable cancers and corresponding synoptic factors (such as primary site, morphology, etc.) is a laborious and time consuming process. Cancer Registries strive to provide timely and accurate information on cancer incidence and mortality in the community. They receive large quantities of data from a range of sources, including hospitals, pathology laboratories and Registries of Births, Deaths and Marriages (which issues releases of death certificates). It is estimated that incident cases within Cancer Registries that have death certificate only notifications amount to about 1-5% of the total cases; delays in the processing of this data may cause underestimation of the incidence of cancer. Computational methods for the automatic abstraction of relevant information have the possibility to enhance a Cancer Registry's workflow, providing time and costs savings as well as timely cancer incidence information and mortality information. This automatic process is however challenging, both for the complex nature of the language used in the reports, and for the high level of recall and accuracy required.

Previous works have attempted to provide automatic cancer coding from free-text pathology reports collected by Cancer Registries. For example, Nguyen et al. [1] used natural language processing techniques and a rule-based system to automatically extract relevant synoptic factors from electronic pathology reports. Similarly, Zuccon et al. [2] showed how these techniques could cope with character recognition errors generated by scanning free-text pathology reports stored in paper form. Machine learning approaches have also been considered; for instance, D'Avolio et al. [3] have tested approaches based on supervised machine learning (Conditional Random fields and Maximum Entropy) and have shown its effectiveness for the classification of pathology reports that were consistent with cancer in the domains of colorectal, prostate, and lung cancer.

Cancer Registries have access to a number of data sources beyond pathology reports. One such data source is death certificates. Death certificates are a rich source of data that can support cancer surveillance, monitoring and reporting. These certificates contain free-text sections that report the cause of the death of an individual. An example of the free-text content of a death certificate where the cause of death is a notifiable cancer is given in Figure 1, while Figure 2 is an example of a non-notifiable death certificate.

Limited works have focused on computational methods for automatically classifing death certificates with respect to the cause of death. The SuperMICAR system and its related tools[1] provide a semi-automatic coding of the cause of death in death certificates. The system identifies keywords and expressions from the free-text documents that indicate possible causes of death; this is done through the use of a standard set of expressions encoded in a predefined vocabulary. Extracted free-text expressions are then converted to one or more

[1] Consult http://www.cdc.gov/nchs/nvss/mmds/super_micar.htm (last visited 19th November 2012) for further details.

```
(I)A) MAXILLARY TUMOR, 2 YEARS B) PULMONARY OEDEMA, 1 WEEK
(II) CEREBROVASCULAR ACCIDENT/DYSPLASIA, 20 YEARS ASTHMA
```

Fig. 1. A de-identified death certificate where the cause of death is a notifiable cancer.

```
I(A) CEREBROVASCULAR ACCIDENT 48 HOURS (B) CEREBRAL ARTERIOSCLEROSIS YEARS
(C) HYPERTENSION YEARS II CHRONIC ALCOHOLISM YEARS
```

Fig. 2. A de-identified death certificate where the cause of death is not a notifiable cancer.

ICD-10 codes which are then aggregated into a single ICD-10 underlying cause of death through the use of a rule-base. While doctor reported death certificates can be fed directly into the system, Coroner reported ones require additional pre-processing. A consistent number (between 15 and 20 percent according to a US study [4]) of death certificates cannot be coded through SuperMICAR and related tools, and thus require manual coding. A recent work has successfully classified death certificates related to pneumonia and influenza using a natural language processing pipeline and rule-based system [5]. However, to the best of our knowledge, no previous research has been conducted to investigate fully automatic methods that go beyond keyword spotting of standard cause of death expressions to classifying death certificates, in particular focusing on certificates where the main cause of death is cancer. Furthermore, while Australian Cancer Registries can acquire free-text death certificates on a fortnightly basis from the Registry of Births Deaths and Marriages, coded causes of death produced by SuperMICAR (and related products) are released by the Australian Bureau of Statistics on a yearly basis. Computational methods able to tackle the fast identification of death certificates where the cause of death is a notifiable cancer would enhance the cancer reporting and monitoring capabilities of Cancer Registries.

In this paper, we focus on the problem of automatically identifying death certificates where the main cause of death is cancer. This problem is cast into a binary classification problem, i.e. death certificates are classified as containing a death cause related to cancer or vice versa as not containing a death cause related to cancer. Several machine learning classifiers were investigated for this task. These include support vector machine, Naive Bayes, decision trees, and boosting algorithms. A state-of-the-art information extraction tool (Medtex [6]) is used to create different set of features that are used to train the classifiers; different feature weighting schemas were also considered. Features include stemmed tokens, n-grams, as well as SNOMED CT concept ids and tokens from fully specified names of SNOMED CT concepts, among others. SNOMED CT is a medical terminology which formally describes in detail the coverage and knowledge of topics and terminology used in the medical domain.

Our approaches are tested on 5,000 de-identified death certificates acquired from an Australian Cancer Registry, using 10-fold cross validation for allowing robust training and testing. Our experimental results demonstrate that the choice of classifier and weighting schema, although being important, is not critical for achieving high classification effectiveness. Instead, the choice of features used to represent content of death certificates is the determining factor for high classification effectiveness. Specifically, stemmed token bigrams are found to be the single most important features among those extracted. Furthermore, we found that SNOMED CT features provide consistent increments in classification effectiveness if used along with stemmed token bigrams; although not providing a large increment, the combined use of stemmed token bigrams and SNOMED CT morphology provide the best classification effectiveness in our experiments.

Next, we detail the approaches adopted in this paper. Then, in Section 3 we outline our empirical evaluation methodology; classification results obtained by the investigated approaches are reported in Section 4. An analysis of the results is developed in Section 4.1. The paper concludes in Section 5 summarising our main contribution and directions for future work.

2 Approaches for Automatic Classification of Death Certificates

In this paper we investigate supervised machine learning approaches for the detection of death certificates where the cause of death is a notifiable cancer. These approaches are characterised by three main variables: (1) the features extracted from the documents (Section 2.1), (2) the weighting schemas applied to the features to represent documents (Section 2.2), and (3) the specific binary classifier used to individuate certificates where the cause of death is a notifiable cancer (Section 2.3).

2.1 Automatic Feature Extraction

Machine learning algorithms require data to be represented by features, such as the words that occur in a text document. We used the information extraction capabilities of the Medtex system[2] for obtaining a set of meaningful features from the free-text of the death certificates.

The feature sets investigated in this paper are:

stem: a token stem, i.e. the stemmed version of a word contained in a certificates

stemBigram: the bi-gram formed by two token stems, i.e. a pair of adjacent stemmed words as found in a certificates

concept: SNOMED CT concepts as found in the free-text of the certificates using the Medtex system

conceptFull: the tokens of the fully specified name of the extracted SNOMED CT concepts

[2] Medtex comprises both information extraction capabilities (extracting both low level information such as word tokens and stems, punctuation, etc., and higher level semantic information such as UMLS and SNOMED CT concepts [1]) and classification capabilities integrated via its rule-based engine.

concFullMorph: the tokens of the fully specified name of extracted SNOMED CT concepts that are morphologic abnormalities or disorders
concBigram: the bigram formed by two adjacent SNOMED CT concept ids
concFullBigram: the bigram formed by two adjacent tokens in the fully specified name of concepts extracted from SNOMED CT

While features like stem and stemBigram are commonly used for classifying free-text documents, features based on SNOMED CT concepts and its properties such as tokens from the fully specified name have not been exploited by previous works that attempted to classify free-text death certificates. SNOMED CT provides a standard clinical terminology used to map various descriptions of a clinical concept to a single standard clinical concept. In this work, the SNOMED CT ontology was used as an underlying mechanism to classify free-text using semantically matching SNOMED CT concepts.

In addition, we also considered pair-wise combinations of features that provided promising results on preliminary experiments. In this paper we shall report the results obtained by all features used singularly, and of the combinations concept + stem, concept + stemBigram, concFullMorph + stemBigram, and concBigram + stemBigram, which has shown promise in preliminary investigations.

Next, we consider the example death certificates given in Figure 1 and Figure 2 to describe how a feature set is constructed. To build the feature representations, we examine each death certificate and for each occurring instance of a feature in the certificate we assign a value of 1, while the absence of a feature is marked by a zero entry value. Note that these values are subsequently modified according to the feature weighting functions, as we shall describe in Section 2.2. After all certificates have been processed in this manner, we add a final feature cancerNotifiable, whose value is obtained from ground truth judgements supplied with the data. Table 1 shows an extract of the feature data constructed for the two example death certificates. The task of the machine learning classifiers is to predict the value of the cancerNotifiable feature, given the learning data supplied.

	Features																			
	stem						stemBigram					concept			conceptFull				...	
Document	ACCID	ALCOHOL	⋮	TUMOR	WEEK	YEAR	ACCID_DYSPLASIA	ACCID_48	⋮	20_YEAR	YEAR_ASTHMA	126550004	⋮	230690007	Neoplasm of maxilla	⋮	Cerebrovascular accident	Cerebral arteriosclerosis	...	cancerNotifiable
Figure 1	1	0	0	1	1	1	1	0	0	1	1	1	0	1	1	1	1	0	...	1
Figure 2	1	1	1	0	0	1	0	1	1	0	0	0	1	1	0	1	1	1	...	0

Table 1. Feature data built from two example death certificates.

Note that no further processing is applied to the text, for example, for removing punctuation, identifying section or list labels, or for removing or correcting typographical errors present in the free-text. While adequate text pre-processing may enhance the quality of the text itself and thus of the extracted features, we left this for future work and instead we focused on investigating weighting schemas for the selected features and binary classifiers.

2.2 Feature Weighting

A number of weighting schemes for capturing the local importance of a feature in a report were tested.

Binary coefficients were used to encode the presence or absence of a feature. We refer to this schema as binary.

The weighting schema composed by the feature frequency $f(\mathcal{F})$ of feature $\mathcal{F}$ was used to capture the number of times a specific feature appeared within a document. We shall refer to this weighting schema as frequency.

Variations of the frequency weighting schema were also experimented with. In this weighting schema, features frequencies were directly translated into weights, i.e. weights are linearly derived from frequencies. Variations consider non-linear functions of the frequency of a feature.

A first variation was to scale the appearance of feature $\mathcal{F}$ in a free-text death certificate by the function $1 + \log(f(\mathcal{F}))$ if $f(\mathcal{F}) \geq 1$, and 0 if the feature was absent. This function would capture the fact that little importance is given to subsequent appearances of a feature $\mathcal{F}$ in a document: the logarithm of a number greater than one plateaus rapidly. In the following, we shall refer to this weighting schema as LogF, i.e. logarithm of the frequency.

A second variation was to assign increasing weights to features that appear with high frequencies within the death certificate. To this aim, the appearance of feature $\mathcal{F}$ was weighted according to the function $e^{f(\mathcal{F})}$, while a zero value was assigned to absent features. It is suggested that, given the short length of the considered death certificates, the unexpected multiple occurrence of a feature would provide strong evidence that that feature is important for the document. Using the exponential function to weight occurrences of a feature would assign dominating scores to features that occur frequently in a document. We shall refer to this weighting function as expF.

Note that only local weighting functions were used to assign scores to features,that is, weights were computed only by taking into account the frequencies of appearance of a feature in a text, thus ignoring the distribution of that feature on a global level, i.e. across the dataset. The incorporation of global occurrence statistics within the weighting schemas is left to future work.

2.3 Automatic Classification Methodology

A number of common classifiers were evaluated. These comprised statistical models (Naive Bayes), support vector machines (SPegasos), decision trees (C4.5), and boosting algorithms (AdaBoost). We considered the implementations of these algorithms provided in the Weka toolkit [7].

The multinomial Naive Bayes classifier determines the class of a death certificate according to the features that occur in the text and their weights. The SPegasos classifier uses a stochastic gradient descent algorithm and a hinge loss function to produce the separation hyperplane used by the linear support vector machine. In the C4.5 classifier, information gain is used for choosing at each level of the decision tree the most effective feature able to split the data into the two binary classes considered here (i.e. death certificates related to cancers and those not related to cancer). Adaboost minimises of a convex loss function built from the prediction of a base weak classifier. A simple binary decision tree classifier that constructs one-level trees was used as base classifier for Adaboost.

Parameters of all classifiers were set to the default values described in Witten et al. [7].

3 Experimental Methodology

3.1 Data

A set of 5,000 free-text death certificates was acquired from Cancer Institute NSW, the institutional entity responsible for maintaining the Central Cancer Registry in New South Wales. Ethics approval was granted by the NSW Population & Health Services Research Ethics Committee for this study including to use the de-identified data. The free-text documents were short in length, containing on average 13.08 words; the (unstemmed) vocabulary contained 3,751 unique words (including section headings and labels).

Cause of death classifications based on ICD-10 codes accompanied the reports. This coding set was acquired from the Australian Bureau of Statistics, who releases coded data yearly. ICD-10 codings were used to determine the class each death certificates belonged to. A list of ICD-10 codes that are cancer notifiable was provided by Cancer Institute NSW.

The 5,000 death certificates were extracted from Cancer Institute NSW archives so that documents were uniformly split across the two classes, i.e. 2,500 certificates were coded with ICD-10 codes that are for notifiable cancers according to the business rules of Cancer Institute NSW, while the remaining 2,500 were not cancer notifiable. The causes of death of the 2,500 death certificates for notifiable cancers span a total of 367 unique ICD-10 codes.

3.2 Evaluation

A 10-fold cross validation methodology was used to train and test the classification algorithms. In this methodology, the dataset was randomly divided into 10 stratified[3] folds of equal dimensions. A model for each classifier was then learnt

[3] Folds were automatically stratified with respect to the two target classes, not the ICD-10 codes.

on nine of these folds, leaving one fold out for testing. The process was repeated by selecting a new fold for testing, while a new model was learnt from the remaining folds. Classification effectiveness was then averaged across the folds left out for testing in each iteration.

F-Measure (F-m) was used as primary metric to evaluate the efficacy of the implemented classifiers; accuracy, recall (sensitivity, Rec) and precision (positive predictive value, Prec) were also recorded, along with the number of true positive (TP), false positve (FP), true negative (TN), and false negative (FN) classifications.

4 Results and Discussion

The combination of 10 features, 4 weighting schemas, and 4 classifiers requires the evaluation of a total of 160 classifier settings (referred to as runs in the following) on the dataset consisting of 5,000 death certificates. While we evaluated all combinations of features, weighting schema and classifiers, given the large number of combinations, it is not feasible to report the individual results for each of the runs. Thus, we report only the settings of the 40 most effective runs in terms on F-measure, our primary evaluation metric (Table 2), with the F-measure of each classifier over all experimented settings graphically shown in Figure 3. Later in the paper we shall consider a summary evaluation of the variability of results provided by features, weighting schemas, and classifiers. This analysis will comprise of the results from all runs.

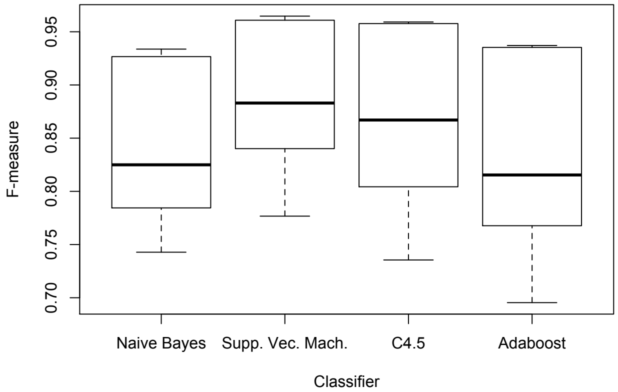

Fig. 3. Boxplot summarising the F-measure performance of the investigated classifiers over all considered settings.

The results reported in Table 2 suggest that the tested approaches are highly effective in discriminating between those death certificates that contain a cancer notifiable cause of death and those that do not.

Classifier	**Feature**	**Weight**	**Prec**	**Rec**	**F-m**	**TP**	**FN**	**FP**	**TN**
SPegasos	concFullMorph + stemBigram	frequency	.9794	.9504	**.9647**	2376	124	50	2450
SPegasos	concFullMorph + stemBigram	logF	.9786	.9500	.9641	2375	125	52	2448
SPegasos	concept + stemBigram	logF	.9770	**.9508**	.9637	**2377**	**123**	56	2444
SPegasos	concFullMorph + stemBigram	binary	.9770	.9504	.9635	2376	124	56	2444
SPegasos	concept + stemBigram	binary	.9766	.9504	.9633	2376	124	57	2443
SPegasos	concept + stemBigram	frequency	.9766	.9504	.9633	2376	124	57	2443
SPegasos	stemBigram	binary	.9761	.9488	.9623	2372	128	58	2442
SPegasos	concFullMorph + stemBigram	expF	.9773	.9476	.9622	2369	131	55	2445
SPegasos	stemBigram	logF	.9753	.9476	.9612	2369	131	60	2440
SPegasos	stemBigram	expF	.9785	.9444	.9611	2361	139	52	2448
SPegasos	stemBigram	frequency	.9764	.9452	.9606	2363	137	57	2443
SPegasos	concept + stemBigram	expF	.9741	.9460	.9598	2365	135	63	2437
C4.5	concept + stemBigram	logF	.9800	.9392	.9592	2348	152	48	2452
C4.5	concept + stemBigram	expF	.9800	.9392	.9592	2348	152	48	2452
C4.5	concept + stemBigram	frequency	.9800	.9392	.9592	2348	152	48	2452
C4.5	concept + stemBigram	binary	.9799	.9384	.9587	2346	154	48	2452
C4.5	concFullMorph + stemBigram	logF	.9856	.9324	.9583	2331	169	34	2466
C4.5	concFullMorph + stemBigram	expF	.9856	.9324	.9583	2331	169	34	2466
C4.5	concFullMorph + stemBigram	frequency	.9856	.9324	.9583	2331	169	34	2466
C4.5	stemBigram	logF	.9848	.9320	.9577	2330	170	36	2464
C4.5	stemBigram	expF	.9848	.9320	.9577	2330	170	36	2464
C4.5	stemBigram	frequency	.9848	.9320	.9577	2330	170	36	2464
C4.5	concFullMorph + stemBigram	binary	.9848	.9320	.9577	2330	170	36	2464
C4.5	stemBigram	binary	.9848	.9308	.9570	2327	173	36	2464
AdaBoost	concept + stemBigram	binary	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concept + stemBigram	logF	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concept + stemBigram	expF	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concept + stemBigram	frequency	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concFullMorph + stemBigram	binary	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concFullMorph + stemBigram	logF	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concFullMorph + stemBigram	expF	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concFullMorph + stemBigram	frequency	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	stemBigram	binary	**1**	.8784	.9353	2196	304	**0**	**2500**
AdaBoost	stemBigram	logF	**1**	.8784	.9353	2196	304	**0**	**2500**
AdaBoost	stemBigram	expF	**1**	.8784	.9353	2196	304	**0**	**2500**
AdaBoost	stemBigram	frequency	**1**	.8784	.9353	2196	304	**0**	**2500**
SPegasos	stem	logF	.9588	.9120	.9348	2280	220	98	2402
SPegasos	stem	frequency	.9611	.9096	.9346	2274	226	92	2408
Naive Bayes	stemBigram	binary	.9658	.9036	.9337	2259	241	80	2420
Naive Bayes	concept + stemBigram	binary	.9606	.9076	.9334	2269	231	93	2407

Table 2. Top 40 results with respect to decrease F-measure (F-m).

Overall, the best classifier is the support vector machine implementation provided by SPegasos when used on concFullMorph + stemBigram features, i.e. the fully specified names of concepts associated to morphological abnormalities and disorders as encoded in SNOMED CT, weighted using raw frequencies. SPegasos is found to be very effective also when other combinations of weighting schemas and features are considered. In addition, this support vector machine classifier shows the smallest variance across all considered settings (Figure 3).

Among the best performing classifiers, AdaBoost used in conjunction with stemmed bigrams features achieved perfect precision (Prec= 1), at the expense of recall. Although these results are remarkable, high precision may be considered less important than high recall in such task. In fact, in a Cancer Registry setting, it is preferable to have high recall and be considering death certificate that are incorrectly reported as containing cancer notifiable cause of death, than to have missed cancer cases. This becomes particularly important if the missed cancer cases refer to rare cancers. AdaBoost also exhibits the highest variance across experiment settings among the considered classifiers (see Figure 3).

4.1 The Impact of Classifiers, Weighting Schemas, and Features

To better understand the role of specific features, weighting schema, and classifiers on the effectiveness of the tested approaches, an analysis of the empirical results where each of the three key characteristics were treated as the controlled variable is performed.

We start by examining the impact of each classification model on the overall effectiveness of the approaches. Table 3 reports maximum (*Max(F-m)*), minimum (*Min(F-m)*), difference (Δ), and variance of F-measure over all runs of each classifier model. SPegasos is found to be the classifier achieving the highest maximum and minimum F-measure values, thus extending the observations made on this classifier when examining the results of Table 2. Instead, while the Naive Bayes classifier was not found to be amongst the most effective classification models in our experiments, its robustness is second only to that of SPegasos, with performance ranges between 0.9337 and 0.7428 in F-Measure. While models such as C4.5 and Adaboost achieve higher values of F-measure than Naive Bayes, their minimum performances are lower than that recorded for Naive Bayes.

Classifier	**Max(F-m)**	**Min(F-m)**	$\mathbf{\Delta}$	**Variance**
SPegasos	**0.9647**	**0.7767**	**0.1880**	$\mathbf{5.10 \cdot 10^{-3}}$
Naive Bayes	0.9337	0.7428	0.1909	$\mathbf{5.10 \cdot 10^{-3}}$
C4.5	0.9592	0.7355	0.2237	$7.35 \cdot 10^{-3}$
AdaBoostM1	0.9371	0.6954	0.2417	$7.88 \cdot 10^{-3}$

Table 3. Classification effectiveness across the four classifiers ordered by increasing max-min F-measure range (Δ).

We continue by analysing the influence of weighting schemas on the classification results of the approaches investigated in this work. Simple raw frequency weighting, i.e. frequency, is found to be the most effective weighting schema. However, no weighting schema appears to be significantly better than another: while

Weight	Max(F-m)	Min(F-m)	Δ	Variance
binary	0.9635	0.6954	0.2681	$6.81 \cdot 10^{-3}$
frequency	**0.9647**	0.6954	0.2693	$6.74 \cdot 10^{-3}$
logF	0.9641	0.6954	0.2687	$6.80 \cdot 10^{-3}$
expF	0.9622	0.6954	**0.2668**	$\mathbf{6.53 \cdot 10^{-3}}$

Table 4. Classification effectiveness across the four weighting schema ordered by increasing max-min F-measure range (Δ).

frequency achieves the best performance with a F-measure of 0.9647, the highest F-measure of the worst performing schema is 0.9622 (expF), just 0.003% lower than frequency. Furthermore, all weighting schema exhibit the same effectiveness when considering the worst performing settings. Thus the range of performance differences and their variance do not significantly differ across weighting schema. This may be due to the fact that death certificates are in general short documents, where features occur uniformly.

Feature	Max(F-m)	Min(F-m)	Δ	Variance
stemBigram	0.9623	**0.9275**	**0.0348**	$\mathbf{2.02 \cdot 10^{-4}}$
concept + bigramStem	0.9637	0.9267	0.0370	$2.16 \cdot 10^{-4}$
concFullMorph + stemBigram	**0.9647**	0.9255	0.0392	$2.33 \cdot 10^{-4}$
concBigram + stemBigram	0.8443	0.7677	0.0766	$8.01 \cdot 10^{-4}$
concBigram	0.8443	0.7677	0.0766	$8.01 \cdot 10^{-4}$
concFullBigram	0.7768	0.6954	0.0814	$8.93 \cdot 10^{-4}$
conceptFull	0.809	0.7177	0.0913	$1.17 \cdot 10^{-3}$
concept + stemBigram	0.9302	0.838	0.0922	$8.39 \cdot 10^{-4}$
concept	0.8743	0.7792	0.0951	$1.13 \cdot 10^{-3}$
stem	0.9348	0.8131	0.1217	$1.36 \cdot 10^{-3}$

Table 5. Classification effectiveness across the ten features ordered by increasing max-min F-measure range (Δ).

Feature is the final variable of our analysis, and the one with the greatest impact on classification results. The use of the concFullMorph + stemBigram feature provide the highest F-measure (0.9647), while concFullBigram yields the lowest maximal F-measure (0.7768): a significant difference of 19.48%. The smallest variance was demonstrated by stemBigram ($2.02 \cdot 10^{-4}$), making it the most robust feature in our experiment; in addition this feature yielded a maximal F-measure of only 0.003% lower than the best value recorded in our experiments. The minimal F-measure yield by the stemBigram feature was also greater than the greatest F-measure values obtained when using half of the features investigated in our study. These results provide strong indication that, of the variables analysed, the choice of feature provides the greatest contribution to the classification effectiveness.

5 Conclusions

Timely processing of cancer notifications is critical for timely reporting of cancer incidence and mortality. Death certificates are a rich source of data on cancer mortality. Cancer registries acquire free-text death certificates on a regular (e.g. fortnightly) basis. However, the cause of death information needs to be classified to facilitate reporting of cancer mortality. Cause of death information classified using ICD-10 codes is only available on an annual basis. In this paper we investigated the automatic classification of death certificates to individuate cancer notifiable cause of deaths. The investigated approaches achieved overall strong classification effectiveness, with a support vector machine classifier trained with token bigram features and information from the SNOMED CT medical ontology, and weighted by their frequency in the documents yielding an F-measure of 0.9647. The choice of features, rather than that of classifiers or weighting schema, was found to be the determining factor for high effectiveness.

Future efforts will be directed towards an in depth error analysis, in particular examining the distance between the prediction produced by a classifier and the decision threshold. We also plan to extend the investigation to predict the actual ICD-10 codes associated to cause of death related to cancer, so as to further assist clinical coders in processing cancer notifications.

References

1. Nguyen, A., Moore, J., Lawley, M., Hansen, D., Colquist, S.: Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications. In: Health Informatics Conference. (2011) 117–124
2. Zuccon, G., Nguyen, A., Bergheim, A., Wickman, S., Grayson, N.: The impact of OCR accuracy on automated cancer classification of pathology reports. Studies in health technology and informatics **178** (2012) 250
3. D'Avolio, L., Nguyen, T., Farwell, W., Chen, Y., Fitzmeyer, F., Harris, O., Fiore, L.: Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). Journal of the American Medical Informatics Association **17**(4) (2010) 375–382
4. Harris, K.: Selected data editing procedures in an automated multiple cause of death coding system. In: Proceedings of the Conference of European Statistics. (1999)
5. Davis, K., Staes, C., Duncan, J., Igo, S., Facelli, J.: Identification of pneumonia and influenza deaths using the death certificate pipeline. BMC Medical Informatics and Decision Making **12**(1) (2012) 37
6. Nguyen, A.N., Lawley, M.J., Hansen, D.P., Bowman, R.V., Clarke, B.E., Duhig, E.E., Colquist, S.: Symbolic rule-based classification of lung cancer stages from free-text pathology reports. Journal of the American Medical Informatics Association **17**(4) (2010) 440–445
7. Witten, I., Frank, E., Hall, M.: Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2011)